2020 2nd International Conference on Artificial Intelligence and Advanced Manufacture (AIAM) | 978-1-7281-9986-3/20/$31.00 ©2020 IEEE | DOI: 10.1109/AIAM50918.2020.00030

Research on Named Entity Recognition for Information Extraction

Qi Guo
Key Laboratory of China's Ethnic Languages and Intelligent Processing of Gansu Province
Northwest Minzu University
Lanzhou, China
e-mail: 1054543194@qq.com

Shuang Wang
Key Laboratory of China's Ethnic Languages and Intelligent Processing of Gansu Province
Northwest Minzu University
Lanzhou, China
e-mail: 1282134950@qq.com

Fucheng Wan*
China's Ethnic Information Technology Research Institute
Northwest Minzu University
Lanzhou, China
e-mail: 306261663@qq.com

***Abstract*—In natural language processing, named entity recognition, as a basic task used in natural language processing, aims at identifying entities with specific meaning in the text, mainly including personal names, place names, institutional names, proper nouns, etc. Named entity recognition technology is an indispensable part of many natural language processing methods, such as information extraction, information retrieval, machine translation and question answering systems. This paper first briefly introduces the developing process of named entity recognition, and describes the basic concept, objectives and difficulties of named entity recognition. Then, it summarizes the methods of named entity recognition from based on rules and dictionary method, based on statistical method, to the deep learning method and migration of learning, and looks forward to the future development of named entity recognition.**

Keywords—named entity recognition, natural language processing, deep learning, transfer learning

I. Introduction

Named entity recognition identifies specific type of named entity from text automatically. Named entities were first proposed at the 1995 MUC-6 conference [1], which specified that entities are unique identifiers and three categories (entity class, time class, and number class) and seven subcategories (person, place, institution, time, date, currency, and percentage) entities to be identified for named entity identification [2]. In the specific application, the actual meaning of the entities are determined the different fields. For example, in the field of entity recognition for electronic medical records [3], we need to think of symptoms and signs, diseases and diagnoses, examinations and tests as named entities; When agricultural pests and diseases are taken as research objects [4], pests, crops and cycles may be taken as named entities. Named entity recognition was first mentioned in the MUC-6 meeting, with the study of named entity recognition, it has been introduced into various evaluation tasks such as ACE, CoNLL, etc. and proposed methods based on traditional machine learning, such as hidden Markov model (HMM), maximum entropy model (ME), conditional random field (CRF), support vector machine (SVM) and so on.

In the early stage, named entity recognition was mainly based on English, because English only needed to consider the characteristics of the word and did not involve word segmentation. The boundary of English words was determined by the blank space, which was relatively easy to realize and obtained good evaluation results. Identify and mark named entities such as names of people, places and organizations in the text, as shown in Fig. 1. There are many difficulties in the process of Chinese named entity recognition [5]: The boundary of Chinese words is not obvious and clear, and there is no display boundary character similar to space; The structure of Chinese named entity is complex; Chinese named entity does not exist morphological transformation and so on. Wu et al. [6] proposed a Chinese emergency extraction method based on enhanced feedback of named entity recognition task, Fb-latiice-Bilstm-CRF, to obtain semantic features of words in sentences, and the F1 value reached 78.80%, with an increase of 5.95%. In order to improve the recognition rate of Chinese named entities, Yao et al. [7] proposed that through integrating the word vector features into the language model XLnet (Generalized autoregressive pretraining for language understanding), the hidden information inside sentences could be fully mined, with F1 value up to 95.73%. Wen et al. [8], aiming at the problem of Chinese named entity recognition in online medical consultation, used the basic framework of pre-trained bidirectional language model and mask language model for transfer learning to improve the performance of Chinese recognition.

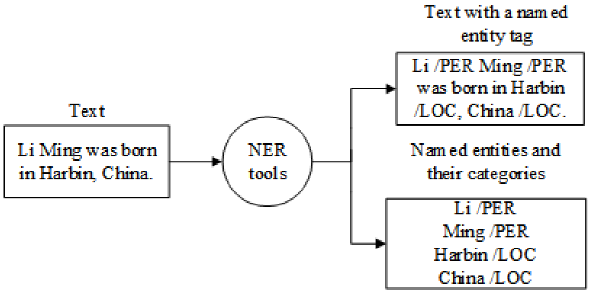
Text with a named entity tag
Li /PER Ming /PER was born in Harbin /LOC, China /LOC.
Text
Li Ming was born in Harbin, China.
NER tools
Named entities and their categories
Li /PER
Ming /PER
Harbin /LOC
China /LOC

Fig. 1. Process of named entity recognition.

II. Traditional Named Entity Recognition Method

In the early stage, named entity recognition was based on manual rules and dictionary library, but it was not feasible due to waste of manpower and material resources, so machine learning method was adopted instead. The statistical machine learning method improves the

978-1-7281-9986-3/20/$31.00 ©2020 IEEE
DOI 10.1109/AIAM50918.2020.00030

performance of the model by selecting appropriate models, methods and feature representation. This method relies on the corpus heavily, so the model is generally trained by combining the two methods. In recent years, with the development of hardware capabilities and the emergence of word embedding, neural network has become a model capable of handling many NLP tasks effectively. The main models are CNN-CRF, RNN-CRF and LSTM-CRF. The method based on deep learning relies on a large number of annotated training data with the same distribution, and the model has poor portability. However, in the practical application, the data are usually small and personalized, so it is very difficult to collect enough training data. Migration learning is introduced in the named entity recognition, and the target domain task model is constructed by using the source domain data and model, in order to improve the amount of annotated data in the target domain and reduce the amount of annotated data required by the target domain model. It has a good effect in dealing with resource-poor named entity recognition tasks. The evolution of named entity recognition is shown in Fig. 2.

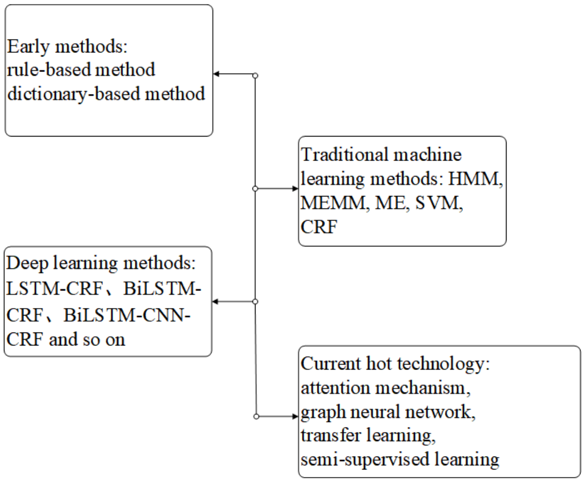

Fig. 2. Development of named entity recognition.

A. Based on Rule and Dictionary Method

The earliest method used in named entity recognition is rule-based method that relies on rule templates constructed manually by linguistics experts and identifies various types of named entities through rule matching. Although the rule-based method can identify the specific field well, it requires linguists to reconstruct the rule templates in different fields. The compilation process is time-consuming, difficult to cover all the rules, and relies on manual operation, which is prone to errors. Most of systems depend on the establishment of knowledge base and dictionary. The system construction cycle is long and the scalability and portability are poor.

B. Based on Statistical Machine Learning Method

Since the late 1990s, statistical methods based on large-scale corpus have gradually become the mainstream of natural language processing, and a large number of machine learning methods have been successfully applied to all aspects of natural language processing. Methods based on statistical machine learning mainly include hidden markov mode (HMM), maxmium entropy (ME), support vector machine (SVM), conditional random fields (CRF), etc. These are shown in TABLE I. This method depends on corpus, but there are few large-scale corpus that can be applied in different fields. Therefore, some scholars improved the model in the later period. Liu et al. [9] proposed a method based on Hidden Semi-Markov Model (HSMM) to realize the prediction which is about the fault of the hydraulic system caused by the mining environment of the coal miner, which is damp and dusty. On the basis of the original K-SVM like model, Tan et al. [10] used grey correlation clustering, complex correlation coefficient method and others to acquire an improved K-SVM multi-classification algorithm, which solved a series of problems such as information overlap among feature variables, high model complexity and low classification accuracy. In the Tibetan named entity recognition based on weak supervised learning, Sun constructed the word representation features to represent the semantic information of the word through the distribution representation of unmarked text words, and added it to the statistical machine learning model of Tibetan name recognition, which improved the recognition effect of the model.

TABLE I. COMPARISON OF NER MODELS

Model	Benefits	Drawbacks
Maximum Entropy Model (ME)	Good versatility	Low training efficiency
Maximum Entropy Markov Model(MEMM)	Make the most of features	Local optimum
Hidden Markov Model (HMM)	Fast training speed	Local optimum
Support Vector Machines (SVM)	Theoretical completeness	Low training efficiency
Conditional Random Fields(CRF)	Flexible features and global optimal	Dependent feature template

C. Hybrid Method

Merge based on rules and dictionary and based on statistical method with Classification techniques like Voting, XVoting, GradingVa, l Grading, etc. The manual rules [11] are introduced into machine learning to train the model and get better recognition effect. Chen et al. proposed a Chinese named entity recognition system that integrates statistics and rules. The Conditional Random Fields (CRF) was used for part-of-speech tagging, and use priority rules to calibrate and filter CRFs model results. In the study of complex named entities, zhou combined statistics and rule methods to conduct extraction experiments on book name, song name and movie name respectively, and reached a better accuracy rate.

III. BASED ON DEEP LEARNING METHOD

A. Advantages of Deep Learning

With the development of deep neural network technology, it has been widely used in named entity recognition. Named entity recognition take advantage of the nonlinear feature of deep learning. In contrast to the linear HMM and CRF, complex features can be obtained from the original data through activation functions. Instead of using a feature-based methods to build features, deep learning can automatically extract information from the input data and learn the representations of the information. Based on the deep learning model is end-to-end, and be able to avoid the error propagation.

B. Deep Learning Structure

- Stage 1: Distributed representations for input. The meaning of a word is determined by the context the word often appear. In other words, words are represented by low-dimensional dense real value vectors, where each dimension represents an implicit feature, so the word can be represented as a continuous dense vector of fixed length. The deep named entity recognition model mainly uses three kinds of distributed representation, they are: word level representation, character level representation and mixed information representation. Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.

- Stage 2: Context encoder. Using input representation to learn semantic encoder. Semantic dependency is obtained through the network, convolutional neural network, recurrent neural network, recursive neural network, transformer and other networks, and potentially endue the models with corresponding prior knowledge.

- Stage 3: Tag decoder. After getting the vector representation of words and converting them into context-dependent representations, the tag decoding module takes them as input and predicts the corresponding tag sequence for the input of the entire model. The mainstream tag decoding structures include: multilayer perceptron and softmax, conditional random fields, recurrent neural network, and pointer Networks.

- Named entity recognition system based on deep learning structure as shown in Fig. 3.

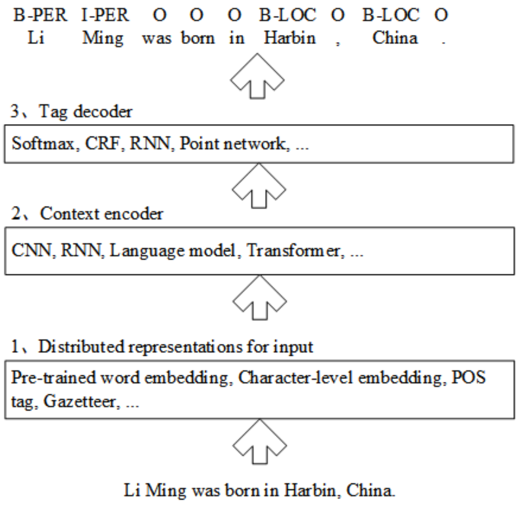
B-PER I-PER O O O B-LOC O B-LOC O
Li Ming was born in Harbin , China .
3、Tag decoder
Softmax, CRF, RNN, Point network, ...
2、Context encoder
CNN, RNN, Language model, Transformer, ...
1、Distributed representations for input
Pre-trained word embedding, Character-level embedding, POS tag, Gazetteer, ...
Li Ming was born in Harbin, China.

Fig. 3. Structure of named entity recognition system based on deep learning

C. Deep Learning Application

Cao et al. used convolutional neural network to deal with sequence labeling problem, used conditional random fields to modify the classification results of the network, and combined the two algorithms to identify the entities in medical records effectively. Gao et al. constructed a mixed model of long short-term memory (LSTM) and conditional random fields (CRF) based on the marked medical cases of lung cancer diagnosis and treatment by famous old Chinese medicine doctors, and used multiple classification evaluation indexes to evaluate the extraction of named entities of symptoms in the medical cases of Chinese medicine. Gao et al. proposed an entity recognition method based on BiLSTM-CRF to identify military named entities such as weapons and equipment, facility targets and troop designation in military texts. As for the study on the recognition of named entities in ancient Chinese, Cui et al. transmitted character sequence information and word sequence information to the Lattice LSTM model. Compared with the BiLSTM-CRF model, its F1 scores increase by about 3.95%. On the basis of constructing Bi LSTM-CRF neural network model, Ma et al. improve the named entity recognition effect of competitive intelligence by adding attention mechanism. Gated recurrent unit (GRU) has one less gate than long short-term memory network, simpler structure and faster training speed. Yu et al. used the CNN-Bidirectional GRU-CRF model to learn features automatically and use context information to realize Chinese word segmentation effectively. On the basis of CNN - BiGRU - CRF model, He et al. integrated attention mechanism to identify alien Marine organisms. Zhou et al. excavated textual data about patient safety incidents for valuable information by combining Chinese semantic and character characteristics to mark the patient safety incident corpus. Cao introduced BERT into BLSTM-CRF model to improve the accuracy of Chinese named entity recognition task due to solve the problem of simplification of word vector representation.

IV. Named Entity Recognition Combining Transfer Learning

A. Transfer Learning

Deep learning depends on the training data strongly, it is able to find the rule in the data through training a large amount of data. In the specific application, Due to the data is often small and personalized [12], the existing data and model cannot be applied to special fields, which caused the poor portability of model. Transfer learning links knowledge or patterns learned in data-rich domains to different but related domains, uses the source domain data and model to construct the task model of the target domain, improves the amount of annotated data in the target domain and reduces the demand of the target domain model on the amount of annotated data, in order to improve the recognition effect of named entities in the special domains. The transfer learning process is shown in Fig. 4.

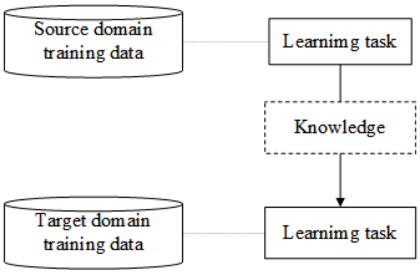

Fig. 4. Transfer learning process

B. Deep Transfer Learning

Deep transfer learning studies how to take advantage of knowledge in other fields through deep neural networks. With the extensive application of deep neural networks in various fields, a large number of deep transfer learning methods have been proposed, which are mainly divided into four categories: instances-based, mapping-based, network-based, and adversarial-based, as shown in Table II. The above techniques are often combined in practical applications. Wang et al. proposed a Chinese named entity recognition mode, Trans-NER, to solve the problem of insufficient data and poor performance with the help of migration learning through extracting text context characteristics, physical characteristics, and adjacent label dependencies to obtain the output result. According to the different correlation between the target data and gives different migration ability, Wang et al. proposed an instance-based transfer learning algorithm, TLNER_AdaBoost, to build a better classifier. Wu et al. solved the problem of insufficient training data and excessive reliance on context meaning in Chinese named entity recognition, and the accuracy rate reached 91.57%, achieving good results.

TABLE II. DEEP TRANSFER LEARNING CLASSIFICATION

Approach Category	**Brief description**	**Method**
Instances-based	Adopt a specific weight adjustment strategy.	TrAdaBoost, TaskTrAdaBoost, BIW
Mapping-based	Map the instance from the source and target domains to the new data space.	CNN, JMMD, Wasserstein instance
Network-based	Part of the pre-trained network in the source domain is reused and transformed into the network for the target domain.	LeNet, AlexNet, VGG, Inception, ResNet
Adversarial-based	The adversarial technique is introduced to find the suitable transportable expression for source and target domain.	Distinguish between the main learning tasks, instead of source and target domains.

V. CONCLUSION

Named entity recognition is the basic work of information extraction, question answering system, machine translation and other tasks. To test entity boundary and named entity type is the foundation of text meaning to understand. The research on named entity recognition has evolved from rule-based and dictionary-based methods to traditional machine learning methods and in recent years to the popular deep learning. At the same time, transfer learning is introduced, which solves the problem of insufficient training data by bridging abundant resources and scarce resources. In addition, we can also study named entity recognition from font, vocabulary information. The application of new technologies, such as antitransfer learning, remote supervised learning, graph neural network and attention mechanism, to named entity recognition will be the focus of future research.

ACKNOWLEDGMENT

During this period of time, the teacher's (Corresponding author: Fucheng Wan) encouragement gave me the motivation to move forward. The complex mood is like every time I do an experiment and I have a result. It is uncontrollable and expectant.

REFERENCES

[1] L. Liu, D. B. Wang, “Overview of Named Entity Recognition”. *Journal of Information Science*, vol. 37, no. 03, pp. 329-340, 2018.

[2] Archana Goyal, Vishal Gupta, Manish Kumar, “Recent Named Entity Recognition and Classification techniques: A systematic review”. vol. 29, pp. 21-43, 2018

[3] C. Chen, X. Y. Liu, Y. H. Fang, “Electronic Medical Record Named Entity Recognition Combining Attention Mechanism”. *Computer Technology and Development*, vol. 30, no. 10, pp. 216-220, 2020,

[4] X. C. Guo, Z. Tang, L. Diao, H. Zhou, L. Li, “Named Entity Recognition of Pests and Diseases Based on Radical Insertion and Attention Mechanism”. *Journal of Agricultural Machinery*, vol. 51, no. S2, pp. 335-343, 2020

[5] G. Wang, “Chinese Entity Recognition and Relationship Extraction Based on Deep Learning”. *Wuhan Institute of Posts and Telecommunications Science*, 2020.

[6] G. L. Wu, J. N. Xu, “Chinese Emergency Extraction Method Based on Named Entity Recognition Task Feedback Enhancement”. Computer Application, pp. 1-8, 2020.

[7] G. B. Yao, Q. G. Zhang, “Chinese Named Entity Recognition Based on XLnet Language Model”. *Computer Engineering and Applications,* pp. 1-9, 2020.

[8] G. H. Wen, H. H. Chen, H. H. Li, et al, “Cross domains adversarial learning for Chinese named entity recognition for online medical consultation”, p. 112, 2012.

[9] X. B. Liu, Y. G. Sun, T. Li, Z. Q. Liu, “Fault Prediction of Hidden Semi-Markov Model for Coal Shearer Adjusting Pump”. *Science, Technology and Engineering*, vol. 20, no. 29, pp. 11980-11986, 2020

[10] X. Tan, G. M. Deng, “K - SVM Classification Algorithm Based on GRC-MCC”. *Statistics and Decision*, no. 22, pp. 10-14, 2012.

[11] F. C. Wan, “Extracting Algorithm for the Optimum Solution Answer Oriented Towards the Restricted Domain”. *Quarterly Journal of Indian Pulp and Paper Technical Association*, p. 12, 2018.

[12] F. C. Wan, “Medical Information Extraction Technology Based on Association Rules”. Indian Journal of Pharmaceutical Sciences, p. 3, 2018.